



Orthogonal Matching Pursuit under the Restricted Isometry Property *

Albert Cohen, Wolfgang Dahmen, Ronald Devore

► To cite this version:

Albert Cohen, Wolfgang Dahmen, Ronald Devore. Orthogonal Matching Pursuit under the Restricted Isometry Property *. Constructive Approximation, 2017, 45, pp.113-127. 10.1007/s00365-016-9338-2 . hal-01239767

HAL Id: hal-01239767

<https://hal.science/hal-01239767>

Submitted on 8 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Orthogonal Matching Pursuit under the Restricted Isometry Property *

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore

December 8, 2015

Abstract

This paper is concerned with the performance of Orthogonal Matching Pursuit (OMP) algorithms applied to a dictionary \mathcal{D} in a Hilbert space \mathcal{H} . Given an element $f \in \mathcal{H}$, OMP generates a sequence of approximations f_n , $n = 1, 2, \dots$, each of which is a linear combination of n dictionary elements chosen by a greedy criterion. It is studied whether the approximations f_n are in some sense comparable to *best n -term approximation* from the dictionary. One important result related to this question is a theorem of Zhang [14] in the context of sparse recovery of finite dimensional signals. This theorem shows that OMP exactly recovers n -sparse signals with at most An iterations, provided the dictionary \mathcal{D} satisfies a Restricted Isometry Property (RIP) of order An for some constant A , and that the procedure is also stable in ℓ^2 under measurement noise. The main contribution of the present paper is to give a structurally simpler proof of Zhang's theorem, formulated in the general context of n -term approximation from a dictionary in arbitrary Hilbert spaces \mathcal{H} . Namely, it is shown that OMP generates near best n -term approximations under a similar RIP condition.

AMS Subject Classification: 94A12, 94A15, 68P30, 41A46, 15A52

Key Words: Orthogonal matching pursuit, best n -term approximation, instance optimality, restricted isometry property.

1 Introduction

Approximation by sparse linear combinations of elements from a fixed redundant family is a frequently employed technique in signal processing and other application domains. We consider such problems in a separable Hilbert space \mathcal{H} endowed with a norm $\|\cdot\| := \|\cdot\|_{\mathcal{H}}$ induced by the scalar product $\langle \cdot, \cdot \rangle$ on $\mathcal{H} \times \mathcal{H}$. A countable collection $\mathcal{D} = \{\varphi_{\gamma}\}_{\gamma \in \Gamma} \subset \mathcal{H}$ is

*This research was supported by the ONR Contracts N00014-11-1-0712, N00014-12-1-0561, N00014-15-1-2181; the NSF Grants DMS 1222715, DMS 0915231, DMS 1222390; the Institut Universitaire de France; the ERC Adv grant BREAD; the DFG SFB-Transregio 40; the DFG Research Group 1779; the Excellence Initiative of the German Federal and State Governments, and RWTH Aachen Distinguished Professorship, Graduate School AICES.

called a *dictionary* if it is complete, i.e., the set of finite linear combinations of elements of the dictionary is dense in \mathcal{H} . The simplest example of a dictionary is the set of elements of a fixed basis of \mathcal{H} . But our primary interest is in *redundant* families. In such a case, there exists a strict subset of \mathcal{D} that is still a dictionary. A primary example of a redundant dictionary is a frame, e.g., any union of a finite number of bases. Without loss of generality we shall always assume that the dictionary \mathcal{D} is *normalized*, i.e.,

$$\|\varphi_\gamma\| = 1, \quad \gamma \in \Gamma.$$

Given such a dictionary \mathcal{D} , we consider the class

$$\Sigma_n = \Sigma_n(\mathcal{D}) := \left\{ \sum_{\gamma \in S} c_\gamma \varphi_\gamma : \#(S) \leq n \right\} \subset \mathcal{H}, \quad n \geq 1. \quad (1.1)$$

The elements in Σ_n are said to be sparse with *sparsity* n . We define

$$\sigma_n(f)_\mathcal{H} := \inf_{g \in \Sigma_n} \|f - g\|,$$

which is called *the error of best n -term approximation* to f from the dictionary \mathcal{D} .

An important distinction between n -term dictionary approximation and other forms of approximation, such as approximation from an n dimensional space, is that the set Σ_n is not a linear space since the sum of two elements in Σ_n is generally not in Σ_n , although it is in Σ_{2n} . Thus n -term approximation from a dictionary is an important example of nonlinear approximation [3] that reaches into numerous application areas such as adaptive PDE solvers, image encoding, or statistical learning. It also serves as a performance benchmark in *compressed sensing* that better captures the robustness of compressed sensing than results on exact sparse recovery (see [2]).

While there are many themes in n -term dictionary approximation, our interest here is in analyzing the performance of greedy algorithms for generating n -term approximations to a given target element $f \in \mathcal{H}$. There are numerous papers on this subject. We refer the reader to the survey article [9] as a general reference. Our particular interest is in understanding what properties of the dictionary \mathcal{D} guarantee that these algorithms perform similarly to best n -term approximation.

These algorithms and best n -term approximation have a simple description when the dictionary \mathcal{D} is an orthonormal or, more generally, a Riesz basis of \mathcal{H} . In this case, the best n -term approximations to a given $f \in \mathcal{H}$ are realized by expanding f in terms of the basis as

$$f = \sum_{\gamma \in \Gamma} c_\gamma \varphi_\gamma \quad (1.2)$$

and retaining n terms from this expansion which correspond to the largest (in absolute value) coefficients. The typical greedy algorithm will construct the same approximations. The situation is much less clear when dealing with more general dictionaries.

In the case of general dictionaries, algorithms for generating n -term approximations are typically built on some form of greedy selection

$$\varphi_k := \varphi_{\gamma_k}, \quad k = 1, 2, \dots, \quad (1.3)$$

of elements from \mathcal{D} and then using a linear combination of $\varphi_1, \dots, \varphi_n$ as the n -term approximation. The standard greedy algorithm (called the Pure Greedy Algorithm) makes the initial selection φ_1 as any element such that

$$\varphi_1 := \operatorname{Argmax}_{\varphi \in \mathcal{D}} |\langle f, \varphi \rangle|. \quad (1.4)$$

This gives the approximation $f_1 := \langle f, \varphi_1 \rangle \varphi_1$ to f and the residual $r_1 := f - f_1$. Given that $\varphi_1, \dots, \varphi_{k-1}$ have been selected, and an approximation f_{k-1} from $F_{k-1} := \operatorname{span}\{\varphi_1, \dots, \varphi_{k-1}\}$ has been constructed, the next dictionary element φ_k is chosen as the best match of the residual

$$r_{k-1} := f - f_{k-1}, \quad (1.5)$$

in the sense that

$$\varphi_k := \operatorname{Argmax}_{\gamma \in \Gamma} |\langle r_{k-1}, \varphi_\gamma \rangle|. \quad (1.6)$$

There exist different ways of forming the next approximation f_k resulting in different greedy algorithms. We focus our attention on *Orthogonal Matching Pursuit* (OMP), which forms the new approximation as

$$f_k := P_k f, \quad (1.7)$$

where P_k is the orthogonal projector onto F_k . OMP is also called the Orthogonal Greedy Algorithm. More generally, we analyze the *Weak Orthogonal Matching Pursuit* (WOMP) where the choice of φ_k is only required to satisfy

$$|\langle r_{k-1}, \varphi_k \rangle| \geq \kappa \max_{\gamma \in \Gamma} |\langle r_{k-1}, \varphi_\gamma \rangle|, \quad (1.8)$$

where $\kappa \in]0, 1]$ is a fixed parameter, which is a more easily implemented selection rule in practical applications. Once this choice of $\varphi_1, \dots, \varphi_k$ is made, then f_k is again defined as the orthogonal projection onto F_k . Notice that the WOMP algorithm includes OMP when one chooses $\kappa = 1$.

The main interest of the present paper is to understand what properties of a dictionary \mathcal{D} guarantee that the approximation rate of WOMP after $O(n)$ steps is comparable to the the best n -term approximation error $\sigma_n(f)$, at least for a certain range $n \leq N$. A related question, but less demanding, is to understand when WOMP is guaranteed to exactly recover f whenever $f \in \Sigma_n$ in $O(n)$ steps for a suitable range of n . This is sometimes referred to as *sparse recovery*. It is known that both of these questions have a positive answer for the entire range of n whenever \mathcal{D} is a Riesz basis for \mathcal{H} (see Corollary 1.1 of [8]).

To give a precise formulation of the type of performance we seek, we define the concept of *instance optimality*.

Instance Optimality: *We say that the WOMP algorithm satisfies instance optimality of order N , if there are constants $A, C > 0$, with A an integer, such that for each f the outputs f_n of WOMP satisfy*

$$\|f - f_{An}\| \leq C \sigma_n(f)_{\mathcal{H}}, \quad (1.9)$$

for $n \leq N$ with C an absolute constant.

Dictionaries for which WOMP is instance optimal in the above sense are called *greedy dictionaries* in [10]. Notice that if (1.9) is satisfied then it implies a positive solution to the sparse recovery problem for the same range of n since $\sigma_n(f) = 0$ when f is in Σ_n .

The subjects of sparse recovery and instance optimality have a long history documented in [10] (see Chapter 2, §2.6). We mention briefly some of the results which serve to orient the present paper. First note that in order to obtain favorable results requires extra structure on the dictionary \mathcal{D} . The main question to be answered is what is the weakest assumptions on the dictionary under which sparse recovery or instance optimality hold. The first positive results were obtained under assumptions on the *coherence* of a dictionary $\mathcal{D} \subset \mathcal{H}$ defined by

$$\mu = \mu(\mathcal{D}) := \sup\{|\langle \varphi, \psi \rangle| : \varphi, \psi \in \mathcal{D}, \varphi \neq \psi\}.$$

For sparse recovery, Tropp [12] proved that whenever the dictionary has coherence $\mu < \frac{1}{2n-1}$, then n steps of OMP recover any $f \in \Sigma_n$ exactly.

For instance optimality, Livschitz [7] proved that whenever $\mu \leq \frac{1}{20n}$, then after $2n$ steps, the OMP algorithm returns $f_{2n} \in \Sigma_{2n}$ such that

$$\|f - f_{2n}\| \leq 3\sigma_n(f)_{\mathcal{H}}. \quad (1.10)$$

A weaker assumption on a dictionary, known as the *Restricted Isometry Property* (RIP), was introduced in the context of compressed sensing [1]. To formulate this property, we introduce the notation

$$\Phi \mathbf{c} = \sum_{\gamma \in \Gamma} c_{\gamma} \varphi_{\gamma}, \quad (1.11)$$

whenever $\mathbf{c} = (c_{\gamma})_{\gamma \in \Gamma}$ is a finitely supported sequence. The dictionary \mathcal{D} is said to satisfy the RIP of order $n \in \mathbb{N}$ with constant $0 < \delta < 1$ provided

$$(1 - \delta)\|\mathbf{c}\|_{\ell^2}^2 \leq \|\Phi \mathbf{c}\|^2 \leq (1 + \delta)\|\mathbf{c}\|_{\ell^2}^2, \quad \|\mathbf{c}\|_{\ell^0} := \#(\text{supp } \mathbf{c}) \leq n. \quad (1.12)$$

Hence this property quantifies the deviation of any subset of the dictionary of cardinality at most n from an orthonormal set. We denote by δ_n the minimal value of δ for which this property holds and remark that trivially $\delta_n \leq \delta_{n+1}$. It is well-known that a coherence bound

$$\mu(\mathcal{D}) < (n - 1)^{-1} \quad (1.13)$$

implies the validity of $\text{RIP}(n)$ for $\delta_n \leq (n - 1)\mu$, but not vice versa [12].

In [14], Tong Zhang proved that OMP exactly recovers finite dimensional n -sparse signals, whenever the dictionary \mathcal{D} satisfies a Restricted Isometry Property (RIP) of order An for some constant A , and that the procedure is also stable in ℓ^2 under measurement noise. There is a simple way to extend Zhang's theorem to more general settings. Our interest is in the following result on instance optimality for WOMP in a general Hilbert space.

Theorem 1.1 *Given the weakness parameter $\kappa \leq 1$, there exist fixed constants A, C, δ^* , such that the following holds for all $n \geq 0$: if \mathcal{D} is a dictionary in a Hilbert space \mathcal{H} for*

which $\text{RIP}((A+1)n)$ holds with $\delta_{(A+1)n} \leq \delta^*$, then, for any target function $f \in \mathcal{H}$, the WOMP algorithm returns after An steps an approximation f_{An} to f that satisfies

$$\|f - f_{An}\| \leq C\sigma_n(f)_{\mathcal{H}}. \quad (1.14)$$

The values of A , C , κ , and δ^* for which the above result holds are coupled. For example, it is possible to have a smaller value of A at the price of a larger value of C or of a smaller value of δ^* . Similarly, a smaller weakness parameter κ can be compensated by increasing A .

The two main differences between Zhang's original theorem and Theorem 1.1 is the treatment of WOMP in place of OMP and the formulation for dictionaries in a Hilbert space rather than the original compressed sensing formulation. Theorem 1.1 can be derived from the original results of Zhang (see [8]) by interpreting the error of best n -term approximation as a measurement noise. In this way, one version of the above result can be derived from [14] for OMP ($\kappa = 1$) with $\delta^* = \frac{1}{3}$ and $A = 30$.

There have been several followups to Zhang's result [8, 11, 5, 6, 13]. These have been directed at either improving the formulation of the result (for example better constants), extending the result to a more general setting, such as Banach spaces, or simplifying the proof. For example, [6], gives the same proof as Zhang, but with different constants $\delta^* = \frac{1}{6}$ and $A = 12$. The paper [5] gives an extension to cover the case of the WOMP. Regarding Theorem 1.1, similar results on instance optimality have been derived in [8] and [11], for the so-called weak Chebyshev greedy algorithm in the more general Banach space setting. In the case that the Banach space is a Hilbert space these results coincide with the WOMP of Theorem 1.1. All of these extensions and modifications utilize the same structure of proof as in the original proof of Zhang.

The main purpose of the present paper is to provide a conceptually more elementary proof for Theorem 1.1. Namely, the proof in [14] and [6] is based on an induction argument which involves an additional inner auxiliary greedy algorithm at each iteration (initialized from a non trivial sparse approximation in each inner loop). Our proof avoids using this auxiliary step. We give the new proof in the following section. We then give some observations that can be derived from Theorem 1.1. We note that our proof is not fighting to obtain the best known constants in Zhang's theorem or its generalization Theorem 1.1.

In this paper, we shall sometimes use the notation $\Phi^*v = (\langle v, \varphi_\gamma \rangle)_{\gamma \in \Gamma}$ for any $v \in \mathcal{H}$, and \mathbf{c}_T to denote, for any $\mathbf{c} = (c_\gamma)_{\gamma \in \Gamma}$ and $T \subset \Gamma$, the sequence whose entries coincide with those of \mathbf{c} on T and are 0 otherwise.

Remark 1.2 *Our results are formulated in the framework of an arbitrary Hilbert space \mathcal{H} , which is in general infinite dimensional, and the dictionary \mathcal{D} has countably infinite cardinality. This raises the question of the practical implementation of OMP or WOMP, since (1.6) involves the search of a maximum within an infinite sequence. In most applications, this problem can be solved by exploiting estimates on the numerical size of the inner products $|\langle r_{k-1}, \varphi_\gamma \rangle|$, which can be obtained from a-priori estimates for the inner products $|\langle f, \varphi_\gamma \rangle|$ and $|\langle \varphi_\gamma, \varphi_\nu \rangle|$. For example, in the case where the dictionary is a wavelet or Gabor frame, smoothness properties of f imply specific estimates for $|\langle f, \varphi_\gamma \rangle|$. Such a-priori estimates typically allow one to reduce the maximum search in (1.6) to a finite subset of \mathcal{D} .*

Remark 1.3 *The RIP property is usually formulated and proved in a finite dimensional context, typically for certain classes of random matrices [6]. Here, we formulate RIP in a possibly infinite dimensional context. Examples of dictionaries from infinite dimensional Hilbert spaces that satisfy RIP include certain redundant frames. As a simple example, for the space $\mathcal{H} = L^2([0, 1])$, consider the dictionary \mathcal{D} consisting of the non-harmonic trigonometric functions*

$$e_n(t) = \exp(i2\pi a n t), \quad n \in \mathbb{Z}, \quad (1.15)$$

for some fixed $0 < a < 1$. With the notation $a = 1 - \varepsilon$, it is then readily seen that for $n \neq m$ one has the coherence property

$$|\langle e_n, e_m \rangle| = \left| \frac{\exp(i2\pi\varepsilon(m-n)) - 1}{2\pi a(n-m)} \right| \leq \frac{\varepsilon}{1-\varepsilon}. \quad (1.16)$$

which is known to imply RIP of order n with constant $\delta = (n-1)\frac{\varepsilon}{1-\varepsilon} < 1$ for $\varepsilon < \frac{1}{n}$. Note that if \mathcal{D} is a Riesz basis in some Hilbert space, then, after proper normalization of the basis vectors, RIP holds at any order n with a fixed δ that depends on the Riesz constants. Thus, in such a case and if δ is sufficiently small, Theorem 1.1 shows that near-best n -term approximations can be produced by application of OMP or WOMP for arbitrary high values of n . This is an interesting alternative to the natural strategy that consists in retaining the n largest components of f in the Riesz basis, since the computation of these components requires manipulating the dual basis which may not have a simple expression.

The remainder of the paper is organized as follows. Section 2 is devoted to the proof of Theorem 1.1 while we conclude in Section 3 with some consequences of this theorem, in particular, highlighting the relation to instance optimal decoders in compressed sensing.

2 Proof of Theorem 1.1

In this section, we give a proof of Theorem 1.1. We begin with the following elementary lemma which guarantees the existence of near best n -term approximations from a dictionary.

Lemma 2.1 *Let \mathcal{D} be a dictionary in a Hilbert space \mathcal{H} that satisfies $\text{RIP}(2n)$. Then,*

- (i) *the set Σ_n of all n -term linear combinations from \mathcal{D} is closed in \mathcal{H} .*
- (ii) *For each $f \in \mathcal{H}$, $\varepsilon > 0$, there exists a $g \in \Sigma_n$ such that*

$$\|f - g\| \leq (1 + \varepsilon)\sigma_n(f)_{\mathcal{H}}. \quad (2.1)$$

Proof: To prove (i), we let $(g^k)_{k \geq 0}$ be a sequence of elements from Σ_n that converges in \mathcal{H} towards some $g \in \mathcal{H}$. We may write

$$g^k = \Phi \mathbf{c}^k = \sum_{\gamma \in \Gamma} c_{\gamma}^k \varphi_{\gamma}, \quad (2.2)$$

with $\|\mathbf{c}^k\|_{\ell^0} \leq n$. For any $\varepsilon > 0$, there exists K such that

$$\|g^k - g^l\| \leq \varepsilon, \quad k, l \geq K. \quad (2.3)$$

From $\text{RIP}(2n)$, it follows that

$$\|\mathbf{c}^k - \mathbf{c}^l\|_{\ell^2} \leq \frac{\varepsilon}{\sqrt{1 - \delta_{2n}}}, \quad (2.4)$$

which shows that the sequence $(\mathbf{c}^k)_{k \geq 0}$ converges in ℓ^2 to some $\mathbf{c} \in \ell^2$. In particular, we find that

$$\lim_{k \rightarrow +\infty} c_\gamma^k = c_\gamma, \quad \gamma \in \Gamma. \quad (2.5)$$

If $c_\gamma \neq 0$ for more than n values of γ , we find that $\|\mathbf{c}^k\|_{\ell^0} > n$ for k sufficiently large which is a contradiction. It follows that $g = \sum_{\gamma \in \Gamma} c_\gamma \varphi_\gamma \in \Sigma_n$.

To prove (ii), let $g^k \in \Sigma_n$ be such that $\|g_k - f\| \rightarrow \sigma_n(f)_\mathcal{H}$. If $\sigma_n(f) > 0$, then $g = g^k$ will satisfy (ii) if k is sufficiently large. On the other hand, if $\sigma_n(f) = 0$, then $g^k \rightarrow f$, $k \rightarrow \infty$. By (i) $f \in \Sigma_n$ and so we can take $g = f$. \square

2.1 Reduction of the residual

Our starting point in proving Theorem 1.1 is the following lemma from [14] which quantifies the reduction of the residuals generated by the WOMP algorithm under the RIP condition. In what follows, we denote by

$$S_k := \{\gamma_1, \dots, \gamma_k\}, \quad (2.6)$$

the set of indices selected after k steps of WOMP applied to the given target element $f \in \mathcal{H}$, and denote as before the residual by $r_k = f - f_k$.

Lemma 2.2 *Let $(f_k)_{k \geq 1}$ be the sequence of approximations generated by the WOMP algorithm applied to f , and let $g = \Phi \mathbf{z}$ with \mathbf{z} supported on a finite set T . Then, if T is not contained in S_k , one has*

$$\|r_{k+1}\|^2 \leq \|r_k\|^2 - \frac{\kappa^2(1 - \delta)}{\#(T \setminus S_k)} (\|r_k\|^2 - \|f - g\|^2), \quad (2.7)$$

where $\delta := \delta_{\#(T \cup S_k)}$ is the corresponding RIP-constant and $\kappa \in]0, 1[$ is the weakness parameter in the WOMP algorithm.

The above lemma quantifies the reduction of $\|r_k\|$ at each iteration, provided that T is not contained in S_k and that $\|f - g\| \leq \|r_k\|$. Note that in the case when $T \subseteq S_k$, we then have $\|r_k\| \leq \|f - g\|$. For the convenience of the reader, we recall the proof of this lemma at the end of this section.

At this point, we depart from the arguments in [14] with the goal of providing a simpler more transparent argument. An immediate consequence of Lemma 2.2 is the following.

Proposition 2.3 *Assume that for a given $A > 0$ and $\delta^* < 1$, $\text{RIP}((A + 1)n)$ holds with $\delta_{(A+1)n} \leq \delta^*$. If $g = \Phi \mathbf{z}$, where \mathbf{z} is supported on a set T such that $\#(T) \leq n$, then for any non-negative integers (j, m, L) such that $0 < \#(T \setminus S_j) \leq m$ and $j + mL \leq An$, one has*

$$\|r_{j+mL}\|^2 \leq e^{-\kappa^2(1-\delta^*)L} \|r_j\|^2 + \|f - g\|^2. \quad (2.8)$$

Proof: By Lemma 2.2, if $g = \Phi \mathbf{z}$ where \mathbf{z} is supported on a set T such that $\#(T) \leq n$, then for any non-negative integers (j, m, L) such that $\#(T \setminus S_j) \leq m$ and $j + mL \leq An$, one has

$$\begin{aligned} \max\{0, \|r_{j+mL}\|^2 - \|f - g\|^2\} &\leq \left(1 - \kappa^2(1 - \delta^*)/m\right)^{mL} \max\{0, \|r_j\|^2 - \|f - g\|^2\} \\ &\leq e^{-\kappa^2(1 - \delta^*)L} \max\{0, \|r_j\|^2 - \|f - g\|^2\}, \end{aligned}$$

where we have used the fact that $\#(T \setminus S_l) \leq m$ for all $l \geq j$. This gives (2.8) and completes the proof of Proposition 2.3. \square

Proof of Theorem 1.1: We fix f and use the abbreviated notation

$$\sigma_n := \sigma_n(f)_{\mathcal{H}}, \quad n \geq 0. \quad (2.9)$$

We first observe that the assertion of the theorem follows from the following.

Claim: *If $0 \leq k < n$ satisfies*

$$\|r_{Ak}\| \leq 2\sigma_k, \quad (2.10)$$

and is such that $\sigma_n < \frac{\sigma_k}{4}$, then there exists $k < k' \leq n$ such that

$$\|r_{Ak'}\| \leq 2\sigma_{k'}. \quad (2.11)$$

Indeed, assuming that this claim holds, we complete the proof of the Theorem as follows. We let k be the largest integer in $\{0, \dots, n-1\}$ for which $\|r_{Ak}\| \leq 2\sigma_k$. Since $\|r_0\| = \sigma_0 = \|f\|$, such a k exists. Since k is maximal, according to the claim, we must have $\sigma_k \leq 4\sigma_n$ and therefore

$$\|r_{An}\| \leq \|r_{Ak}\| \leq 2\sigma_k \leq 8\sigma_n, \quad (2.12)$$

we see that (1.14) holds with $C = 8$.

We are therefore left with proving the claim. For this, we fix

$$\delta^* = \frac{1}{6}, \quad (2.13)$$

and $0 \leq k < n$ such that (2.10) holds and such that $\sigma_n < \frac{\sigma_k}{4}$. Let $k < K \leq n$ be the first integer such that $\sigma_K < \frac{\sigma_k}{4}$. By (ii) of Lemma 2.1 we know that for any $B > 1$ there is a $g \in \Sigma_K$ with $\|f - g\| \leq B\sigma_K$. Therefore, g has the form

$$g = \Phi \mathbf{z} = \sum_{\gamma \in T} z_\gamma \varphi_\gamma, \quad \#(T) = K. \quad (2.14)$$

The significance of K is that on the one hand

$$\|f - g\| \leq B\sigma_K < \frac{B}{4}\sigma_k, \quad (2.15)$$

while on the other hand

$$\sigma_k \leq 4\sigma_{K-1}. \quad (2.16)$$

To eventually apply Proposition 2.3 for the above g and $j = Ak$, we need to bound $\#(T \setminus S_{Ak})$ with A yet to be specified. To this end, we write $K = k + M$, with $M > 0$, and observe that if $S \subset T$ is any set with $\#(S) = M$ and $g_S := \sum_{\gamma \in S} z_\gamma \varphi_\gamma$, then

$$\|g_S\| \geq \|f - (g - g_S)\| - \|f - g\| \geq \sigma_k - B\sigma_K \geq \left(1 - \frac{B}{4}\right)\sigma_k, \quad (2.17)$$

where we have used the fact that $g - g_S \in \Sigma_k$. Using RIP, we obtain the following lower bound for the coefficients of g : for any set $S \subset T$ of cardinality M , we have

$$\left(1 - \frac{B}{4}\right)^2 \sigma_k^2 \leq \|g_S\|^2 \leq (1 + \delta^*) \sum_{\gamma \in S} |z_\gamma|^2 = \frac{7}{6} \sum_{\gamma \in S} |z_\gamma|^2. \quad (2.18)$$

Next define S_g to be the set of the M smallest coefficients in absolute value of g . Then for any $S \subset T$ with $\#(S) \geq M$, one has $\left(\sum_{\gamma \in S} |z_\gamma|^2\right) / \left(\sum_{\gamma \in S_g} |z_\gamma|^2\right) \geq \#(S)/M$, and hence,

$$\frac{6}{7} \left(1 - \frac{B}{4}\right)^2 \frac{\#(S)}{M} \sigma_k^2 \leq \sum_{\gamma \in S} |z_\gamma|^2. \quad (2.19)$$

Now, we consider the particular set $S := T \setminus S_{Ak}$. For such a set, if we write

$$g - f_{Ak} = \sum_{\gamma \in T \cup S_{Ak}} \tilde{z}_\gamma \varphi_\gamma, \quad (2.20)$$

we find that $\tilde{z}_\gamma = z_\gamma$ when $\gamma \in S$. Therefore, if $\#(S) \geq M$,

$$\frac{6}{7} \left(1 - \frac{B}{4}\right)^2 \frac{\#(S)}{M} \sigma_k^2 \leq \sum_{\gamma \in T \cup S_{Ak}} |\tilde{z}_\gamma|^2. \quad (2.21)$$

and the above bound combined with the RIP implies

$$\begin{aligned} (1 - \delta^*)^{\frac{6}{7}} \left(1 - \frac{B}{4}\right)^2 \frac{\#(S)}{M} \sigma_k^2 &\leq \|g - f_{Ak}\|^2 \leq (\|g - f\| + \|r_{Ak}\|)^2 \\ &\leq (B\sigma_K + 2\sigma_k)^2 \leq \left(\frac{B}{4} + 2\right)^2 \sigma_k^2, \end{aligned}$$

Since $\delta^* = 1/6$ this gives the bound

$$\#(T \setminus S_{Ak}) \leq \frac{7 \left(\frac{B}{4} + 2\right)^2}{5 \left(1 - \frac{B}{4}\right)^2} M \leq 13M, \quad (2.22)$$

where the second inequality is obtained by taking B sufficiently close to 1.

We proceed now verifying the claim with $k' = K - 1$ when $K - 1 > k$ and with $k' = k + 1$ otherwise. In the first case we can use the reduction estimate provided by Proposition 2.3 with $j = Ak$ in combination with (2.16) to deal with the term $\|r_{Ak}\|$ in (2.8). When $K = k + 1$, however, we cannot bound $\|r_{Ak}\|$ directly in terms of a σ_l for some $l > k$. Accordingly, we use Proposition 2.3 in different ways for the two cases.

In the case where $M \geq 2$, i.e., $K - 1 > k$, we apply (2.8) with $j = Ak$, $m = 13M$ and $L = \lceil 4\kappa^{-2} \rceil$. Indeed $Ak + Lm = Ak + 13M\lceil 4\kappa^{-2} \rceil \leq An$ holds for $k + M \leq n$ whenever $A \geq 13\lceil 4\kappa^{-2} \rceil$. Moreover, notice that for such an A

$$A(K - 1) = Ak + A(M - 1) \geq Ak + \frac{1}{2}AM = Ak + \frac{Am}{26} \geq Ak + Lm, \quad (2.23)$$

whenever

$$A \geq 26\lceil 4\kappa^{-2} \rceil. \quad (2.24)$$

This gives

$$\begin{aligned} \|r_{A(K-1)}\|^2 &\leq \|r_{Ak+Lm}\|^2 \\ &\leq e^{-10/3} \|r_{Ak}\|^2 + \|f - g\|^2 \\ &\leq e^{-10/3} 4\sigma_k^2 + B^2 \sigma_K^2 \\ &\leq e^{-10/3} 64\sigma_{K-1}^2 + B^2 \sigma_{K-1}^2 \\ &\leq 4\sigma_{K-1}^2, \end{aligned}$$

where we have used (2.16) in the fourth inequality, and the last inequality follows by taking B sufficiently close to 1. We thus obtain (2.11) for the value $k' = K - 1 > k$.

In the case $M = 1$, i.e., $K = k + 1$, we apply (2.8) with $j = Ak$, $m = 13$ and

$$L = \lceil 6\kappa^{-2} \rceil.$$

In fact, from (2.22) we know that $\#(T \setminus S_{Ak}) \leq 13$ and $An \geq A(k + 1) \geq Ak + mL$ for A satisfying (2.24). This yields

$$\begin{aligned} \|r_{A(k+1)}\|^2 &\leq \|r_{Ak+mL}\|^2 \\ &\leq e^{-5} \|r_{Ak}\|^2 + \|f - g\|^2 \\ &\leq 4e^{-5} \sigma_k^2 + B^2 \sigma_{k+1}^2 \\ &\leq \left(4e^{-5} + \frac{B^2}{16}\right) \sigma_k^2. \end{aligned}$$

This implies that $S_{A(k+1)}$ contains T . Indeed, if it missed one of the indices $\gamma \in T$, then we infer from the RIP,

$$\begin{aligned} (1 - \delta^*)|z_\gamma|^2 &\leq \|g - f_{A(k+1)}\|^2 \\ &\leq (\|f - g\| + \|r_{A(k+1)}\|)^2 \\ &\leq \left(B\sigma_K + \sqrt{4e^{-5} + \frac{B^2}{16}}\sigma_k\right)^2 \\ &\leq \left(\frac{B}{4} + \sqrt{4e^{-5} + \frac{B^2}{16}}\right)^2 \sigma_k^2. \end{aligned}$$

On the other hand, we know from (2.19) that

$$\frac{6}{7} \left(1 - \frac{B}{4}\right)^2 \sigma_k^2 \leq |z_\gamma|^2, \quad (2.25)$$

which for B sufficiently close to 1 is a contradiction since $\frac{6}{7}\left(1-\frac{B}{4}\right)^2 > \frac{6}{5}\left(\frac{B}{4} + \sqrt{4e^{-5} + \frac{B^2}{16}}\right)^2$. This implies that $\|r_{A(k+1)}\| \leq \sigma_{k+1}$, and therefore (2.11) holds for the value $k' = k + 1$. This verifies the claim and hence completes the proof of Theorem 1.1. \square

Proof of Lemma 2.2: First observe that, since $r_k = f - P_k f$ is the orthogonal projection error,

$$\begin{aligned}\|r_{k+1}\|^2 &= \|f - P_{k+1}f\|^2 \\ &= \|f - P_k f\|^2 - \|(P_k - P_{k+1})f\|^2 \\ &\leq \|r_k\|^2 - |\langle r_k, \varphi_{\gamma_{k+1}} \rangle|^2.\end{aligned}$$

Therefore, it suffices to prove that $\|r_k\|^2 - |\langle r_k, \varphi_{\gamma_{k+1}} \rangle|^2$ is bounded by the right hand side of (2.7) which amounts to showing that

$$(1 - \delta)(\|r_k\|^2 - \|f - g\|^2) \leq \kappa^{-2} \#(T \setminus S_k) |\langle r_k, \varphi_{\gamma_{k+1}} \rangle|^2. \quad (2.26)$$

We may assume that $\|r_k\| \geq \|f - g\|$ otherwise there is nothing to prove. To prove (2.26), we first note that

$$\begin{aligned}2\|g - f_k\|\sqrt{\|r_k\|^2 - \|f - g\|^2} &\leq \|g - f_k\|^2 + \|r_k\|^2 - \|f - g\|^2 \\ &= \|g - f_k\|^2 + \|r_k\|^2 - \|g - f_k - r_k\|^2 \\ &\leq 2|\langle g - f_k, r_k \rangle| = 2|\langle g, r_k \rangle|.\end{aligned}$$

This is the same as

$$\|r_k\|^2 - \|f - g\|^2 \leq \frac{|\langle g, r_k \rangle|^2}{\|g - f_k\|^2}. \quad (2.27)$$

If we write $f_k = \Phi \mathbf{c}^k$, with \mathbf{c}^k supported on S_k , then the numerator of the right side satisfies

$$\begin{aligned}|\langle g, r_k \rangle| &= |\langle \Phi \mathbf{z}, r_k \rangle| \\ &= |\langle \mathbf{z}_{S_k^c}, \Phi^* r_k \rangle_{\ell_2}| \\ &\leq \|\mathbf{z}_{S_k^c}\|_{\ell_1} \|\Phi^* r_k\|_{\ell_\infty} \\ &\leq \kappa^{-1} \|\mathbf{z}_{S_k^c}\|_{\ell_1} |\langle r_k, \varphi_{\gamma_{k+1}} \rangle| \\ &\leq \kappa^{-1} \sqrt{\#(T \setminus S_k)} \|\mathbf{z}_{S_k^c}\|_{\ell_2} |\langle r_k, \varphi_{\gamma_{k+1}} \rangle| \\ &\leq \kappa^{-1} \sqrt{\#(T \setminus S_k)} \|\mathbf{z} - \mathbf{c}^k\|_{\ell_2} |\langle r_k, \varphi_{\gamma_{k+1}} \rangle|.\end{aligned}$$

On the other hand, recalling that $\delta = \delta_{\#(S_k \cup T)}$, the denominator satisfies by the RIP,

$$\|g - f_k\|^2 = \|\Phi(\mathbf{z} - \mathbf{c}^k)\|^2 \geq (1 - \delta) \|\mathbf{z} - \mathbf{c}^k\|_{\ell_2}^2. \quad (2.28)$$

Therefore we have obtained

$$\|r_k\|^2 - \|f - g\|^2 \leq \frac{\#(T \setminus S_k) |\langle r_k, \varphi_{\gamma_{k+1}} \rangle|^2}{\kappa^2 (1 - \delta)}, \quad (2.29)$$

which is (2.26). \square

3 Near-best n -term approximations

We close this paper with some remark concerning the type of near-best n -term approximations that can be obtained using WOMP.

Let us observe first that Theorem 1.1 does not give that f_n is a near-best n -term approximation in the form

$$\|f - f_n\| \leq C_0 \sigma_n(f)_{\mathcal{H}}, \quad (3.1)$$

where n is now equal on both sides of the inequality. However, a simple *postprocessing* of f_{An} by retaining its n largest components in absolute value does satisfy (3.1).

Theorem 3.1 *Under the assumptions of Theorem 1.1, let $f_{An} = \Phi \mathbf{c}^{An}$ be the output of WOMP after An steps. Let $T \subset \Gamma$, $\#(T) = n$, be a set of indices corresponding to n largest (in absolute value) entries of \mathbf{c}^{An} . Define $f_n^* \in \Sigma_n$ to be the element obtained by retaining from f_{An} only the n -terms corresponding to the indices in T . Then,*

$$\|f - f_n^*\| \leq C^* \sigma_n(f)_{\mathcal{H}}, \quad (3.2)$$

where the constant C^* depends on the constant C in Theorem 1.1 and on the RIP-constant $\delta_{(A+1)n}$.

Proof: By Lemma 2.1, there exists a \mathbf{c} with $\|\mathbf{c}\|_{\ell_0} \leq n$, such that

$$\|f - \Phi \mathbf{c}\| \leq 2\sigma_n(f)_{\mathcal{H}}. \quad (3.3)$$

It follows that

$$\|\mathbf{c} - \mathbf{c}^{An}\|_{\ell^2} \leq \frac{1}{\sqrt{1 - \delta_{(A+1)n}}} \|\Phi \mathbf{c} - \Phi \mathbf{c}^{An}\|_{\ell^2} \leq \frac{C + 2}{\sqrt{1 - \delta_{(A+1)n}}} \sigma_n(f)_{\mathcal{H}}. \quad (3.4)$$

Since $\|\mathbf{c}^{An} - \mathbf{c}_T^{An}\|_{\ell^2} \leq \|\mathbf{c}^{An} - \mathbf{c}\|_{\ell^2}$, we have

$$\begin{aligned} \|\mathbf{c} - \mathbf{c}_T^{An}\|_{\ell^2} &\leq \|\mathbf{c} - \mathbf{c}^{An}\|_{\ell^2} + \|\mathbf{c}^{An} - \mathbf{c}_T^{An}\|_{\ell^2} \\ &\leq 2\|\mathbf{c} - \mathbf{c}^{An}\|_{\ell^2}, \end{aligned} \quad (3.5)$$

which, by (3.4), provides

$$\|\mathbf{c} - \mathbf{c}_T^{An}\|_{\ell^2} \leq 2\|\mathbf{c} - \mathbf{c}^{An}\|_{\ell^2} \leq \frac{2(C + 2)}{\sqrt{1 - \delta_{(A+1)n}}} \sigma_n(f)_{\mathcal{H}}. \quad (3.6)$$

The approximation $\Phi \mathbf{c}_T^{An}$ is in Σ_n and satisfies

$$\|f - \Phi \mathbf{c}_T^{An}\| \leq 2\sigma_n(f)_{\mathcal{H}} + \|\Phi(\mathbf{c}_T^{An} - \mathbf{c})\| \leq \left(2 + \frac{2\sqrt{1 + \delta_{(A+1)n}}(C + 2)}{\sqrt{1 - \delta_{(A+1)n}}}\right) \sigma_n(f)_{\mathcal{H}}, \quad (3.7)$$

which proves (3.2). \square

Remark 3.2 When asking instead for recovering the signal \mathbf{c} from $f = \Phi \mathbf{c}$ one should note though that the best n -term approximation of \mathbf{c}^{An} provided by WOMP is not necessarily a near-best approximation to the signal \mathbf{c} in ℓ^2 . The recovery of \mathbf{c} from $f = \Phi \mathbf{c}$ up to $\sigma_n(\mathbf{c})_{\ell^2}$, also called instance optimality in ℓ^2 , was discussed in [2] where it is proved that this objective cannot be attained under general RIP conditions. Nevertheless, the results in [2] combined with Zhang's theorem also show that, in the finite-dimensional setting, instance-optimality in ℓ^2 holds for any \mathbf{c} with high probability when the sensing matrices are drawn from standard families of random matrices.

References

- [1] E. Candes, J. Romberg, T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure and Appl. Math., 59 (2006), 1207–1223.
- [2] A. Cohen, W. Dahmen, R. DeVore, *Compressed Sensing and best k -term approximation*, J. Amer. Math. Soc. 22 (2009), 211–231.
- [3] R.A. DeVore, *Nonlinear approximation*, Acta Numerica, 7 (1998), 51–150.
- [4] R. DeVore, G. Petrova, P. Wojtaszczyk, *Instance-optimality in probability with an ℓ_1 -minimization decoder*, Appl. Comput. Harm. Analysis, 27(3)(2009), 275–288.
- [5] S. Foucart, *Stability and robustness of weak orthogonal matching pursuits. Recent Advances in Harmonic Analysis and Applications*, Springer Proceedings in Mathematics and Statistics, 25 (2013), 395–405.
- [6] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser Verlag, Basel, 2013.
- [7] E.D. Livshitz, *On the optimality of the Orthogonal Greedy Algorithm for μ -coherent dictionaries*, Journal of Approximation Theory 164-5 (2012), 668–681.
- [8] E.D. Livshitz and V.N. Temlyakov, *Sparse approximation and recovery by greedy algorithms*, IEEE Trans. Inf. Theory, 60 (2014), 3989–4000
- [9] V.N. Temlyakov, *Greedy approximation*, Acta Numerica, 17 (2008), 235–409.
- [10] V. N. Temlyakov, *Greedy Approximation*, Cambridge University Press, 2011.
- [11] V. N. Temlyakov, *Sparse approximation and recovery by greedy algorithms in Banach spaces*, Forum Math. Sigma, 2 (2014), DOI: <http://dx.doi.org/10.1017/fms.2014.7>.
- [12] J.A. Tropp, *Greedy is good: algorithmic results for sparse approximation*, IEEE Trans. Inf. Theory, 50 (2004), 2231–2242.
- [13] J. Wang and B. Shim, *Exact recovery of sparse signals via orthogonal matching pursuit: How many iterations do we need?*, arXiv:1211.429304, 2015.

- [14] T. Zhang, *Sparse recovery with orthogonal matching pursuit under RIP*, IEEE Trans. Inform. Theory, 57(9) (2011), 6215–6221.

Albert Cohen

Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 75005, Paris, France
cohen@ann.jussieu.fr

Wolfgang Dahmen

Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55,
D-52056 Aachen Germany
dahmen@igpm.rwth-aachen.de

Ronald DeVore

Department of Mathematics, Texas A&M University, College Station, TX 77840, USA
rdevore@math.tamu.edu